데이타마이닝 기법을 이용한 문서 자동 분류 모델

김영인* · 이진용** · 문현정* · 우용태*

An Automatic Text Classification Model using Association Rules

In-Young Kim*, Jin-Yong Lee**, Hyeon-Jeong Mun*, Yong-Tae Woo*

요 약

기업에서 보유한 전문 지식 정보가 급속도로 증가함에 따라 대량의 문서에 저장된 지식 정보를 효과적으로 탐색하여 기업 경영에 활용하기 위한 지식경영시스템 도입이 확산되고 있다. 이러한 지식경영시스템에서 핵심적인 구성 요소는 전문 분야의 지식 정보를 체계적으로 분류하고 효율적으로 검색하기 위한 지식 탐사 기법이다. 본 논문에서는 데이타마이닝 기법을 이용하여 문서를 자동적으로 분류하기 위한 새로운 모델을 제안하였다. 연관 규칙 탐사 알고리즘을 이용하여 학습용 문서 집합으로부터 세부 분야를 대표하는 색인어 집합을 구성하였다. 세부 분야별 색인어 집합에 대하여 전체 문서에 대한 비중에 따라 가중치 배열을 구성하여 문서를 자동으로 분류하기 위한 기준으로 삼았다. 임의의 문서를 자동적으로 분류하는 실험을 통하여 제안된 방법의 효율성을 검정하였다.

Keywords : 데이터마이닝, 문서자동분류, 연관규칙

1. 서론

정보화 사회에서 인간의 지적 활동을 지원하기 위한 필수 불가결한 조건의 하나는 정보의 수집과 활용에 있다. 또한 지식기반 사회가 도래함에 따라 기업에서도 기업이 보유한 전문 지식 정보를 체계적으로 분류하여 기업 경영에 활용하기 위한 지식경영시스템을 활발하게 도입하고 있다. 이러한 지식경영시스템에서 핵심적인 구성 요소는 전문 분야의 지식 정보를 체계적으로 분류하고 효율적으로 검색하기 위한 지식 탐사 기법이다. 하지만 대부분의 기업에서 보유하고 있는 수많은 전문 정보들은 컴퓨터에 의해 디지털화 되고 있지만 지식경영시스템이라 할 수 있는 수준에는 이르지 못하고 있다.

최근에는 데이타마이닝 기법을 이용하여 대량의 데이타로부터 숨겨진 패턴을 분석하여 유용한 정보를 찾기 위한 연구가 활발하게 전개되고 있다[1]. 텍스트마이닝은 데이타마이닝 기법중에서 텍스트 데이타로부터 연관성을 분석하기 위한 접근 방법의 하나이다. 텍스트마이닝은 크게 텍스트 요약, 텍스트 범주화, 텍스트 군집화로 구분된다. 텍스트 요약은 주어진 텍스트의 내용에서 대표가 되는 일부 내용을 사용자에게 제시하여 전체적인 내용을 쉽게 파악할 수 있도록 도와주는 기능이다. 텍스트 범주화는 분야별로 문서를 분류하고 임의의 문서에 대해 해당되는 범주로 자동으로 분류하는 기능이다. 텍스트 군집화는 주어진 텍스트 집합을 내용의 유사성에 따라 소집단으로 분할하는 기능을 말한다[2].

기업에서 보유한 전문 지식 정보가 급속도로 증가함에 따라 이러한 정보를 자동적으로 분류하여 원하는 지식을 효율적으로 검색하기 위한 방법이 필수적으로 요구되고 있다. 일반적으로 문서에 대한 자동 분류 과정은 문서 내용에서 중심이 되는 대표 단어를 찾아내기 위한 색인어 추출 과정과 추출된 색인어를 이용하여 문서간의 연관성에 따라 문서를 분류하는 클러스터링 과정으로 구성된다. 여기서 색인어는 정보 제공자와 이용자를 연결하기 위한 중간매개체 역할을 하는 기능으로 주로 TF*IDF 알고리즘이나 벡터·길이 정규·

* 창원대학교 전자계산학과
**거제대학 전자계산과

화 알고리즘 등과 같은 확률 모델을 이용하여 각 문서를 대표할 수 있는 단어를 추출하여 단어들 간의 연관성을 계산한다[3, 4]. 또한 문서 자동 분류나 정보검색 과정에서 주로 사용되는 클러스터링은 문서에 포함된 단어와 같은 식별 요소를 이용하여 유사한 문서끼리 하나의 클러스터로 구성하기 위한 방법이다.

문서간의 유사도를 측정하여 유사한 문서끼리 하나의 클러스터로 구성하기 위한 기준으로 다양한 형태의 유사계수가 사용되고 있다. 이러한 유사계수는 대상의 특성에 따라 거리계수, 연관계수, 상관계수, 확률적 유사계수 등으로 구분된다[5]. 또한 문서를 자동적으로 분류하기 위하여 주로 사용하는 유사계수는 코사인, 다이스, 자카드 계수 등이 있다[3]. 이러한 유사계수는 문서, 용어 또는 사용자의 질의어 등의 요소에 따라 문서간의 유사도를 측정하기 위해 사용되고 있다. 하지만 유사계수에 따라 분류 결과가 다르게 나타나거나 문서 클러스터에 적합하지 못한 유사계수도 있다. 유사계수는 기본적으로 각 용어가 나타난 문서의 총 개수와 동시 출현한 문서 개수의 비로 정의되어 하나의 문서가 여러 클러스터링에 중복해서 포함되거나 클러스터에 포함되지 않을 수도 있다.

본 논문에서는 전문 지식 정보를 효율적으로 검색하기 위하여 데이타마이닝 기법을 이용하여 전문 분야에 대한 문서를 자동적으로 분류하기 위한 새로운 모델을 제시하였다. 먼저, 세부 분야별로 학습용 문서 집합으로부터 대표 용어를 추출하여 분야별 색인어 집합을 구성하였다. 색인어 집합을 구성하기 위해 특정 분야와 관련된 문서 집합에서 추출한 전문 용어를 대상으로 일정 지지도와 신뢰도를 만족하는 전문 용어들간의 연관 규칙을 발견하였다. 이러한 연관 규칙을 이용하여 각 세부 분야를 대표하는 색인어 집합을 구성하였다. 연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문서에서 추출된 전문 용어 집합이다. 각 트랜잭션을 구성하는 전문 용어 집합은 형태소 분석을 통하여 추출된 모든 용어에 대하여 전문 용어 사전에 수록된 용어만을 추출하여 구성하였다. 각 색인어 집합에 대하여 전체 문서에 대한 비중에 따라 가중치 배열을 구성하였다. 이 가중치 배열은 임의의 문서에서 추출된 용어 집합과 가장 유사한 세부 분류를 찾기 위해 사용된다. 결과적으로 임의의 문서에서 추출된 전문 용어 집합과 가장 유사한 클러스터에 의해 해당 문서를 자동적으로 분류할 수 있다.

제안한 문서 자동 분류 방법의 효율성을 검증하기 위하여 컴퓨터 관련학회에서 발표한 240건의 학술 논문을 대상으로 실험을 하였다. 각 세부 분야에 대한 대표 색인어 집합을 구성하기 위하여 학회에서 분류한 8개의 세부 분야별로 30편씩 논문을 선정하여 논문에서 추출한 전문 용어를 대상으로 연관 규칙 탐사 알고리즘을 적용하였다. 문서 자동 분류 실험은 색인어 구성에 사용되지 않은 논문에서 추출한 전문 용어 집합과 세부 분야별 대표 색인어와 유사도를 비교하여 세부 분야를 찾는 과정으로 이루어진다. 기존의 유사도 계수에 의한 비교 실험을 통하여 본 논문에서 제안한 방법의 효율성을 입증하였다.

2. 문서 자동 분류 기법

컴퓨터 기술의 발전으로 데이타의 양이 기하급수적으로 증가함에 따라 사용자가 원하는 지식 정보를 찾는 과정이 점점 어려워지고 있다. 컴퓨터를 이용하여 문서를 자동으로 분류하기 위한 시도는 70년대 말 Salton에 의해 체계화되기 시작하였다[6]. 최근에는 대량의 문서를 자동적으로 분류하여 효율적으로 지식 정보를 검색하기 위한 문서 자동 분류 시스템에 대한 연구가 활발하게 전개되고 있다. 문서를 자동적으로 분류하기 위한 연구 방법은 사전에 정의된 분류 체계에 따라 문서를 집단화하는 자동 분류 기법, 사전 분류 체계 없이 문서간의 유사성에 의해 문서를 집단화하는 클러스터링 기법, 그리고 연관 규칙을 이용한 클러스터링 방법이 있다[7]. 또한 문서간의 유사도를 측정하기 위한 다양한 형태의 유사 계수가 연구되었다[8].

2.1. 자동 분류를 위한 대표 색인어 추출

문서를 자동적으로 분류하기 위해 여러 가지 방법들을 복합적으로 이용하고 있다. 이중에서 가장 핵심적인 기술은 문서에서 대표 색인어를 추출하기 위한 방법이다. 대표 색인어를 추출하기 위한 방법으로 하나의 문서에 포함된 단어들의 가중치를 조절하는 TF*IDF 알고리즘이 널리 사용되고 있다. 벡터 길이 정규화(Vector Length Normalization) 알고리즘은 문서 길이에 따른 불균형을 해결하기 위한 방법이다. 또한 색인 대상 단어 수를 줄이기 위하여 접두사나 어근에 붙어 있는 부분을 제거하는 어근 추출 알고리즘을 사용한다. 그리고 일반적인 단어를 제거하기 위한 불용어 제거 알고리즘이나 동일한 뜻을 가진 단어를 표준화 하기 위하여 동의어 사전도 이용하고 있다[9].

2.2. 클러스터링 기법에 의한 문서 자동 분류

문서의 자동 분류나 정보 검색 과정에서 주로 사용되는 클러스터링 기법은 문서를 대표하는 색인어를 이용하여 유사한 문서끼리 클러스터를 형성하는 방법이다. 이 기법을 이용하여 문서를 자동적으로 분류하기 위한 연구도 다양하게 진행되고 있다. 김호성 등[7]은 제목에 나타난 단어의 출현 빈도와 단어간의 연관성에 의해 논문을 자동적으로 분류할 수 있는 클러스터링 시스템을 구현하였다. 하지만 논문 제목에서 주제 분야와

관련된 용어가 없거나 단어의 출현 빈도가 적어서 클러스터에 포함되지 않은 경우가 발생할 수 있다. 신진섭 등[9]은 웹 상의 문서를 사용자 프로파일에 따라 분류하기 위한 클러스터링 모델을 제시하였다. 문서의 자동 분류는 각 프로파일에 속한 단어에 대한 가중치와 문서내의 단어의 빈도 수를 곱하여 최대 값을 가지는 프로파일에 따라 문서를 분류한다. 하지만 실험에서 사용한 문서의 수가 적어서 제시한 분류 기법의 신뢰성을 검증하기 어렵다.

2.3. 연관규칙을 이용한 문서 클러스터링

최근에는 데이타마이닝 기법을 이용한 클러스터링에 대한 연구가 진행되고 있다. Han 등[10]은 연관 규칙을 사용하여 최소 지지도를 만족하는 모든 빈발 항목 집합을 대상으로 하이퍼그래프를 생성하였다. 생성된 하이퍼그래프에 의해 신뢰도에 근거한 유사성 척도를 사용하여 분할하는 클러스터링 알고리즘을 제시하였다. 신성보 등[11]은 트랜잭션에 대한 클러스터의 유사성을 측정하기 위해 주요 항목과 비 주요 항목으로 구분하고 각 트랜잭션에 대한 최소 비용 계산을 통해 자동화된 문서 클러스터링 기법을 제안하였다. 생성한 클러스터내에서 연관 규칙을 사용하여 용어들간의 연관성을 발견하였다. 하지만 주요 항목 집합에 대한 기준은 단순히 용어의 출현 빈도 수만 고려하는 관계로 무의미한 연관 규칙이 대량으로 발견될 수 있다.

2.4. 문서간의 유사도를 측정하기 위한 유사계수

클러스터링의 기준이 되는 유사도를 측정하기 위하여 식별요소의 벡터를 이용한 다양한 형태의 유사계수가 사용되고 있다. 유사계수는 거리계수, 연관계수, 상관계수, 확률적계수 등이 있다. 거리계수는 공간상에서의 거리에 따라 대상간의 비유사성을 측정하는 방법이고, 연관계수는 비교 대상이 가지는 속성의 일치 정도를 측정하는 방법이다. 상관계수는 비교 대상을 표현하는 속성들의 벡터 쌍에 대한 독립성을 측정하는 방법이며 확률적 유사계수는 정보량 공식에 기반하여 두 사건의 확률 변수간의 의존관계를 정량적으로 표현한 것이다. 이러한 유사계수중에서 문서 클러스터링에서 자주 쓰이는 유사계수는 연관계수이다. 대표적으로 다이스 계수(Dice Coefficient), 자카드 계수(Jaccard Coefficient), 코사인 계수(Cosine Coefficient) 등이 있다[5].

3. 문서 자동 분류 모델

3.1. 데이타마이닝 기법을 이용한 문서 자동 분류 모델

본 논문에서는 전문 지식 정보를 효율적으로 검색할 수 있는 새로운 형태의 문서 자동 분류 방법을 제안하였다. 제안한 방법은 데이타마이닝 기법중에서 대량의 데이타로부터 숨겨진 패턴을 찾기 위해 사용되는 연관 규칙 탐사 알고리즘을 이용하여 학습용 문서 집합으로부터 대표 용어를 추출하여 세부 분야별 색인어 집합을 구성하였다. 그리고 임의의 문서에서 추출한 용어를 이용하여 가장 유사한 세부 분야로 자동적으로 분류할 수 있는 방법을 제안하였다.

제안한 방법의 문서 분류 과정은 크게 전처리 과정, 학습용 용어로부터 대표 용어 추출을 통한 색인어 구성 과정 그리고 임의의 문서를 자동적으로 분류하는 과정으로 구성된다. 다음 그림 1은 본 논문에서 제안한 문서 자동 분류 시스템의 전체적인 구성도이다.

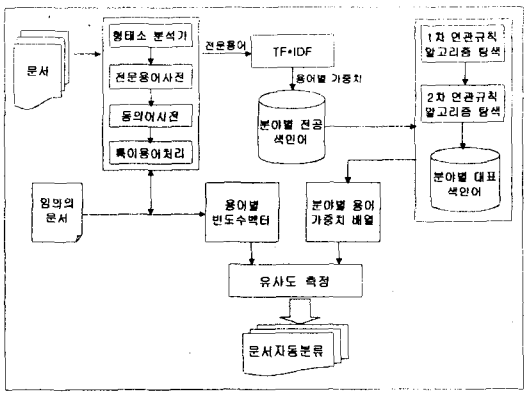

그림 1. 문서 자동 분류 시스템의 전체 구성도

3.2. 전처리 과정

전처리 과정은 색인어 집합을 구성하기 위해 필요한 전문 용어를 학습용 문서로부터 추출하기 위한 과정이다. 본 논문에서는 자동 분류를 위한 실험을 컴퓨터 분야의 논문을 대상으로 실험하였다. 따라서 본 과정은 컴퓨터 분야의 논문에서 컴퓨터 용어만을 별도로 추출하기 위한 과정으로 실험 대상 문서에서 학습용 전문 용어를 추출하기 위한 과정, 동의어 사전 적용 과정, 특이 용어 제거 과정 그리고 단어 빈도 수에 따른 가중치 적용 과정으로 이루어진다. 다음 그림 2는 전처리 과정에 대한 전체적인 구성도이다.

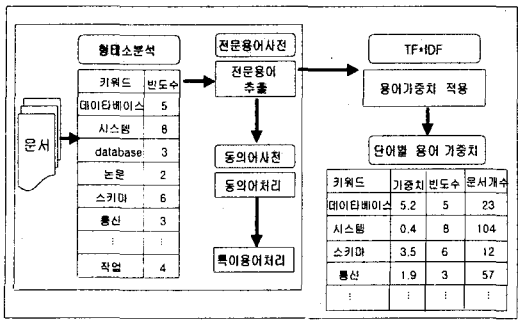

그림 2. 전처리 과정에 대한 전체 구성도

3.2.1. 학습용 전문 용어 추출

먼저, 학습용 문서에 대한 형태소 분석을 통하여 문서에서 출현하는 모든 용어를 추출하였다. 형태소 분석기는 형태소 분석기는 한성대학교 강승식 교수팀이 개발한 공개용 형태소 분석기인 HAM4.0a[12]를 사용하였다. 그리고 학습용 문서에서 추출된 단어에서 신성대학 김현숙 교수의 컴퓨터용어 사전에 수록된 컴퓨터 용어만을 별도로 추출하였다.

3.2.2. 동의어 사전

전문 용어중에서 같은 의미를 가진 용어이지만 저자에 따라 영어나 한국어 용어를 혼용하고 있다. 특히 영어로 된 전문 용어를 한글로 표기하는 경우에서 자주 발생한다. 이러한 동의어는 별도의 동의어 사전을 구성하여 용어를 표준화하였다. 예를 들어 '데이터베이스', '데이타베이스', 'database', 'databases', 'db' 등과 같은 용어는 하나의 용어로 통일하였다.

3.2.3. 특이 용어 처리

전체 문서에서 출현하는 절대 빈도 수가 매우 적은 용어는 연산시간만 낭비하고 최소지지도를 만족하지 못하기 때문에 연관 규칙으로 발견되지 않는다. 그리고 전문 용어이지만 모든 분야에서 공통적으로 사용되는 전문 용어는 특정 분야를 대표하는 용어로 보기 어렵다. 본 논문에서는 이러한 용어를 특이 용어로 처리하여 연관 규칙 탐사 과정에서 제외시켜 무의미한 연관 규칙의 양산을 방지하여 대표 색인어를 효율적으로 추출할 수 있도록 하였다.

3.2.4. 단어 빈도 가중치

문서를 자동적으로 분류하는 과정에서 문서를 대표하는 특징 단어를 추출하기 위하여 단어에 대한 빈도수(Term Frequency)를 가장 많이 이용한다. 하지만 문서에서 출현하는 단어의 빈도 수가 높다고 해서 그 문서를 대표하는 단어가 된다고 확신하기는 어렵다. 예를 들어 '시스템' 이라는 용어는 컴퓨터 용어이지만 대부분의 컴퓨터 관련 논문에서 공통적으로 출현하고 또한 빈도수도 높기 때문에 특정 분야를 대표하는 용어로 판정하기 어렵다. 이러한 단어 빈도수에 의한 문제점을 해결하기 위하여 여러 가지 형태의 가중치 공식들이 제안되었다[13].

본 논문에서는 TF*IDF 알고리즘을 적용하여 모든 문서에서 공통적으로 출현하는 단어에 대한 가중치를 조정하였다. TF*IDF 알고리즘은 하나의 문서에서 출현하는 단어의 빈도수에 대하여 역 문서 빈도수(Inverse Document Frequency)를 가중치로 적용하여 문서를 대표하는 단어들을 효과적으로 선별하기 위한 방법이다. 다음 식 (1)은 TF*IDF 알고리즘에서 특정 단어에 대한 중요도를 계산하는 식이다.

$$w_{ij} = tf_{ij} \log(N/df_i) \qquad (1)$$

식 (1)에서 df_i는 N개의 문서들 중에서 단어 t_j가 존재한 문서의 개수를 의미하며, tf_{ij}는 문서 d_i에서 단어 t_j가 나타난 수를 의미한다. 이때 $\log(N/df_i)$는 역 문서 빈도수를 의미한다. w_{ij}는 역 문서 빈도수와 단어 빈도수를 곱한 값을 문서 d_i에서 단어 t_j의 중요도 또는 영향력(Weight)이 된다.

3.2.5. 문서 길이 정규화

일반적으로 문서들의 길이가 다른 관계로 각 문서에서 추출된 단어들은 문서 길이에 따라 영향력이 차이가 난다. 따라서 문서 길이의 차이에 따른 영향력의 불균형을 해결하기 위해 여러 가지 형태의 벡터 길이 정규화(Vector Length Normalization) 알고리즘을 사용하고 있다[9]. 하지만 본 논문의 실험 대상 문서는 학회 발표 논문을 모집단으로 선정하여 대부분의 문서 길이가 일정하기 때문에 문서 길이 정규화 과정은 생략하였다.

3.3. 분야별 대표 색인어 구성

문서를 자동적으로 분류하기 위해 모 집단을 여러 개의 클러스터로 분할하는 과정이 필요하다. 따라서 문서 자동 분류 과정에서 가장 핵심적인 과정은 각 클러스터를 효율적으로 분할하기 위하여 분야별 기준이 되는 대표 색인어를 효율적으로 구성하는 과정이다.

본 논문에서는 데이타마이닝 기법을 이용하여 특정 분야를 대표하는 색인어 집합을 자동적으로 추출하기 위한 새로운 방법을 제안하였다. 먼저, 연관 규칙 탐사 알고리즘을 이용하여 특정 분야

에서 추출된 전문 용어 간의 연관성을 분석하였다. 연관성 분석은 특정 분야에서 추출된 전문 용어 집합에 대하여 해당 분야와 가장 밀접한 전문 용어 집합을 1차로 추출하였다. 그리고 1차로 추출된 전문 용어 집합의 각 용어에 대하여 연관 규칙 탐사 알고리즘을 재 적용하여 색인어 집합을 확장하였다. 마지막으로 분야별 색인어 집합에 대해 빈도순으로 일정 개수의 용어를 최종적인 대표 색인어 집합으로 선택하였다. 다음 그림 3은 연관 규칙 탐사 알고리즘을 이용하여 분야별 대표 색인어를 추출하는 과정에 대한 개념도이다.

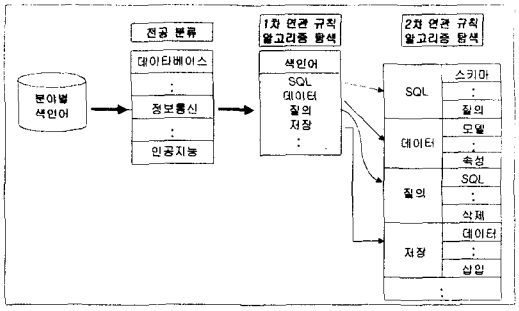

그림 3. 연관 규칙 탐사를 이용한 분야별 대표 색인어 추출

3.3.1. 연관 규칙의 정의

연관 규칙이란 '어떤 사건이 발생하면 다른 사건이 일어난다'와 같은 연관성을 의미한다. 연관 규칙 탐사 알고리즘에서 하나의 장바구니에 담긴 상품 집합이나 단위 시간에 발생한 사건들의 묶음을 트랜잭션이라 정의한다. 연관 규칙 탐사란 이러한 트랜잭션 집합에서 최소 지지도와 신뢰도를 만족하는 의미 있는 연관 규칙을 발견하는 과정을 말한다[14].

다음은 연관 규칙에 대한 정의이다. 먼저, $I = \{1, 2, 3, \cdots, m\}$을 항목들의 집합, D를 트랜잭션들의 집합이라 하면 각 트랜잭션 T는 $T \subseteq I$인 항목들의 집합이다. X를 한 트랜잭션에 포함된 항목들의 빈도를 고려하지 않은 항목들의 집합일 때 $X \subseteq T$ 이면 트랜잭션 T는 X를 포함한다. 이때 연관 규칙(association rule)은 R: $X \rightarrow Y$ 형식의 함축이고, 이때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다. 만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측하는 것이 연관 규칙이다. 이런 확률을 이 규칙에 대한 신뢰도(conf(R))라 한다. R의 신뢰도는 다음 식(2)처럼 X를 지지하는 T에 대하여 Y 또한 지지할 조건부 확률로 정의된다.

$$conf(R) = p(Y \subseteq T \mid X \subseteq T) = \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} = \frac{supp(X \cup Y)}{supp(X)} \qquad (2)$$

D에 있는 규칙 R에 대한 지지도는 $supp(X \cup Y)$로 정의한다. 지지도는 얼마나 자주 적용할 수 있는 지를 나타내는 반면 신뢰도는 그 규칙이 얼마나 믿을만한지를 의미한다. 규칙이 데이타베이스에서 적절해지려면 충분한 지지도와 신뢰도를 가져야 한다. 그러므로 어떤 주어진 최소 신뢰도 c_{min}과 최소지지도 s_{min}에 대하여 $conf(R) \geq c_{min}$ 이고 $supp(R) \geq s_{min}$하면 규칙 R은 D에 대하여 성립한다고 할 수 있다. 규칙이 성립되기 위해 필요한 조건으로서 규칙의 조건부와 결과부는 모두 빈발해야 한다.

3.3.2. 연관 규칙 탐사 알고리즘

연관 규칙을 탐사하기 위한 알고리즘은 다음과 같은 2 단계로 이루어진다.

1 단계 : 먼저, 빈발 항목 집합(large itemsets) l을 찾아낸다. 항목들의 전체집합 I의 부분 집합이면서 몇 개의 항목들로 구성된 것을 항목집합이라 한다. 여기서 최소 지지도 s_{min} 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들을 빈발 항목 집합이라 한다.

2 단계 : 모든 빈발 항목집합 l에 대하여 l의 모든 공집합이 아닌 부분집합들을 찾는다. 이러한 부분집합 a에 대하여 $supp(a)$에 대한 $supp(l)$의 비율이 적어도 최소 신뢰도 c_{min}이상 , 즉 $\frac{supp(l)}{supp(a)} \geq c_{min}, a \Rightarrow (l-a)$ 의 형태의 규칙을 연관 규칙으로 출력한다.

3.3.3. 연관 규칙을 이용한 분야별 색인어 구성

분야별 색인어 집합을 구성하기 위해 연관 규칙 탐사 알고리즘을 이용하여 최소 지지도와 신뢰도를 만족하는 전문 용어들간의 연관 규칙을 발견하였다. 이러한 연관 규칙은 각 분야별로 관련된 전문 용어를 클러스터링하여 문서를 자동으로 분류하기 위한 기준으로 사용된다. 즉, 특정 분야와 관련된 문서 집합에서 추출한 전문 용어를 대상으로 발견된 연관 규칙을 이용하여 각 세부 분야를 대표하는 색인어 집합을 구성하였다.

연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문서에서 추출된 전문 용어 집합이다. 각 트랜잭션을 구성하는 전문 용어 집합은 전처리 과정에서 형태소 분석을 통하여 추출된 모든 용어에 대하여 전문 용어사전에 수록된 용어만을 추출하여 구성하였다. 그리고 같은 의미를 가지는 동의어를 표준화하고 불필요한 연산이나 연관 규칙을 양산할 수 있는 특이 용어도 제외 시켰다.

3.3.4. 전문 용어별 가중치 배열 구성

색인어로 추출된 모든 전문 용어에 대하여 전체 빈도 수에 대한 비율에 의해 분야별로 가중치 배열을 구성하였다. 각 전문 용어에 대한 가중치 W_{ij}는 다음과 같이 정의하였다. 여기서 N_j는 각 전문 용어에 대한 전체 출현 빈도 수이고, D_{ij}는 세부 분야별 출현 빈도 수이다.

$$W_{ij} = D_{ij} / N_j \qquad (3)$$

다음 표 1은 분야별로 가중치 배열을 구성한 예이다.

표 1. 가중치 배열 구성 예

	데이타베이스	소프트웨어공학	통신	인공지능	전산교육	…
스키마	0.8	0.1	0.0	0.0	0.1	…
소켓	0.0	0.2	0.8	0.0	0.0	…
캡슐화	0.1	0.6	0.2	0.1	0.0	…
객체	0.3	0.5	0.1	0.1	0.0	…
인터페이스	0.2	0.1	0.2	0.0	0.5	…
유사성	0.0	0.2	0.0	0.7	0.1	…

가중치 배열에서 각 전문 용어에 대한 가중치는 특정 용어가 해당 세부 분야에서 차지하는 중요도를 의미한다. 즉, 각 세부 분야를 대표하는 색인어 집합에 대한 가중치의 합은 다른 분야의 가중치 합보다 큰 값을 가지게 되어 해당 클러스터를 구분하기 위한 기준이 된다. 이러한 가중치 배열을 이용하여 임의의 문서에서 추출된 전문 용어의 가중치와 비교하여 문서를 자동적으로 분류할 수 있다.

3.4. 가중치 배열을 이용한 문서 자동 분류

임의의 문서에 대한 자동 분류 과정은 다음과 같다. 먼저, 전처리 과정을 통하여 분류 대상 문서에서 전문 용어만을 추출하여 식 (3)에 의해 용어별로 가중치를 계산한다. 그리고 다음 유사도 측정 식 (4)에 의해 분류 대상 문서가 속하는 세부 분야를 자동적으로 분류할 수 있다. 여기서 n은 색인 집합의 전체 크기이고, T_i는 분류 대상 문서에서 추출된 각 전문 용어의 빈도 수이다.

$$Cluster = Max\, C_j \quad (C_j = \sum_{i=1}^{n} (T_i * W_{ij})) \qquad (4)$$

C_j는 분류 대상 문서와 각 세부 분야별 유사도를 의미한다. 결국 모든 세부 분야를 대상으로 계산한 유사도중에서 최대 값을 가지는 분야가 해당 문서가 속하게 될 세부 분야가된다. 다음 그림 4는 각 분야별 용어별 가중치 배열과 임의의 문서에서 추출된 용어에 의한 유사도 계산을 통하여 문서를 자동적으로 분류하는 과정이다.

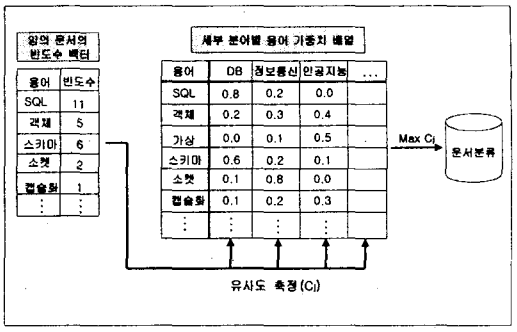

그림 4. 유사도에 의한 임의 문서의 자동 분류

4. 실험 결과 및 고찰

본 논문에서 제안한 방법의 효율성을 검증하기 위하여 컴퓨터 관련 학회에서 발표된 280건의 학술 논문을 대상으로 실험을 하였다. 학회에서 분류한 8개의 세부 분야별로 30편씩 선정하여 분야별 대표 색인어 집합을 구성하였다. 분야별 대표 색인어를 구성하기 전처리 과정을 통해 추출된 전문 용어에 대해 연관 규칙 탐사 알고리즘을 적용하였다. 임의의 문서에 대한 자동 분류 실험은 색인어 구성 과정에서 사용되지 않은 논문을 분야별로 5편씩 선정하여 실험하였다. 기존의 유사도 계수에 의한 비교 실험을 통하여 본 논문에서 제안한 방법의 효율성을 입증하였다.

240건의 학술 논문을 전처리 과정을 거쳐 추출된 전문 용어는 총 1,499건이 추출되었다. 다음 표 2는 세부 분야별 논문에서 전처리 과정을 통해 추출된 전공 용어의 수이다.

표 2. 분야별 전문 용어의 수

순번	세부 분야	전문용어 수
1	데이타베이스	737
2	멀티미디어	780
3	소프트웨어공학	680
4	인공지능	619
5	전산수학 및 교육	655
6	정보보안	755
7	정보통신	689
8	화상 및 음성처리	504

실험 대상 논문에서 추출된 전체 전문 용어는 1,499건이고 멀티미디어 분야가 780건으로 가장 많다. 또한 화상 및 음성처리 분야를 제외한 대부분의 분야에서 600개 이상으로 전체 논문에서 추출된 전문 용어와 40%이상 중복되고 있다. 따라서 전문 용어별 단순 빈도 수에 의해 문서를 자동 분류하기는 어렵고, 각 세부 분야별로 대표 색인어에 대한 가중치를 다르게 부여할 필요가 있다.

분야별 대표 색인어를 구성하는 과정은 다음과 같다. 먼저, 연관 규칙 알고리즘을 이용하여 각 세부 분야내의 전문 용어중에서 가장 연관성이 있는 단어를 추출하였다. 각 세부 분야의 문서를 트랜잭션 단위로 하여 최소 지지도 5%와 최소 신뢰도 50%를 만족하는 연관 규칙을 발견하였다. 그리고 1차 연관 규칙 알고리즘에서 추출된 색인어에 대하여 2차 연관 규칙 알고리즘을 적용시켰다.

다음 표 4는 연관 규칙 알고리즘을 통하여 추출된 데이타베이스 분야의 대표 색인어 집합이다.

표 4. 데이타베이스 분야의 색인어 집합

분야	대표 색인어	대표 색인어 개수
데이타 베이스	데이타베이스, 객체, 질의, 검색, 관계, 테이블, 접근, 디자인, SQL, 표현, 특징, 속성, 언어, 개념, 삭제, 이름, 스키마, 식별자, 컨텐츠, 변경, 참조, 타입, 변환, 인터페이스, select, 지향, 연결, 위치, 공간, 객체지향	30

다음 그림 5는 각 세부 분야에서 추출된 대표 색인어의 분포 현황이다. 그림에서처럼 각 분야에서 주로 사용되는 색인어는 해당 분야에서 높은 분포를 보이는 것을 알 수 있다.

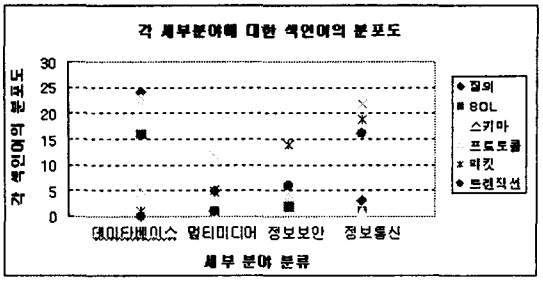

그림 5. 세부 분야의 대표 색인어 분포 현황

표 5는 대표 색인어 구성에 참여하지 않은 데이타베이스 분야의 임의 논문에 대하여 본 논문에서 제안한 유사도 계수를 분야별로 계산한 결과이다. 실험 결과에서처럼 이 논문은 데이터베이스 분야와 가장 연관된 논문임을 알 수 있다. 그리고 이 논문은 데이터베이스 분야 다음으로 관련된 분야는 인공지능 분야이고, 가장 관련성이 낮은 분야는 화상 및 음성처리 분야임을 알 수 있다.

표 5. 각 세부 분야에 따른 임의의 데이타베이스 분야의 유사도

대분류	유사도 값
데이타베이스	316.16296
인공지능	120.96201
소프트웨어공학	94.80456
전산수학 및 교육	82.47659
정보보안	62.25558
멀티미디어 및 HCI	56.59221
정보통신	29.65092
화상 및 음성처리	28.09517

다음 표 6은 색인어 구성에 사용하지않은 40편의 논문을 대상으로 본 논문에서 제안한 유사도 계수와 기존의 유사 계수에 의해 분류 실험을 한 결과이다.

표 6. 제안 모델과 기존 유사 계수간의 자동 분류 실험 결과

유사계수 / 세부 분야	제안 모델	다이스계수	자카드계수	코사인계수
데이타베이스	**4**	0	0	2
멀티미디어	**3**	0	0	0
소프트웨어공학	**5**	4	4	4
인공지능	**2**	2	4	4
전산수학 및 교육	**4**	3	3	4
정보보안	**5**	2	2	3
정보통신	**3**	5	4	3
화상 및 음성처리	**5**	5	5	5
전체 논문	**78%**	40%	55%	62%

표 6에서처럼 본 논문에서 제안한 유사도 계수에 의해 성공적으로 분류한 비율은 78%정도로 다른 3종류의 유사도 계수보다 더 정확하게 분류하였고 3개의 분야에서 100% 성공률을 보였다. 특히 멀티미디어 분야의 경우에서는 기존의 유사계수는 모두 실패하였지만, 본 모델에서는 60%의 성공률을 보였다. 하지만 인공지능 분야의 논문은 본 모델에서도 2편만 성공적으로 분류하였고, 2편은 소프트웨어공학 분야로 그리고 1편은 데이터베이스 분야로 분류되었다.

5. 결론

본 논문에서는 데이타마이닝 기법을 이용하여 전문 분야에 대한 문서를 자동적으로 분류하기 위한 새로운 모델을 제시하였다. 학습용 문서 집합으로부터 대표 용어를 추출하여 분야별 대표 색인어 집합을 구성하였다. 색인어 집합을 구성하기 위해 연관 규칙 탐사 기법을 이용하였다. 분야별 색인어 집합에 대하여 전체 문서에 대한 비율을 이용한 가중치 배열을 구성하여 새로운 형태의 유사도를 정의하였다.

제안한 문서 자동 분류 방법의 효율성을 검증하기 위하여 컴퓨터 관련학회에서 발표한 240건의 학술 논문을 대상으로 실험을 하였다. 각 세부 분야에 대한 대표 색인어 집합을 구성하기 위하여 학회에서 분류한 8개의 세부 분야별로 30편씩 논문을 선정하여 논문에서 추출한 전문 용어를 대상으로 연관 규칙 탐사 알고리즘을 적용하였다.

문서 자동 분류 실험은 색인어 구성에 사용되지 않은 논문에서 추출한 전문 용어 집합과 세부 분야별 대표 색인어간의 유사도를 비교하였다. 실험 결과 본 모델에서 제안한 유사도에 의해 성공적으로 분류한 비율은 78%정도로 다이스 계수, 자카드 계수, 코사인 계수 보다 더 정확하게 분류하였다. 특히 3개의 분야에서 100% 성공률을 보였고, 멀티미디어 분야에서는 기존의 유사 계수는 모두 실패하였지만, 본 모델에서는 60%의 성공률을 보였다.

앞으로 사전 분류과정 없이 세부 분류를 자동적으로 클러스터링 할 수 있는 모델을 개발하여 실용적으로 응용 가능한 지식 탐사 시스템을 개발할 계획이다.

참고문헌

[1] 나민영, "대규모 지식데이타베이스에서 유용한 지식 추출하는 기법," 데이타베이스 월드, pp.5-17, 1997.7

[2] 조태호, "텍스트 마이닝에 대한 소개와 기능," 한국정보처리학회 추계학술논문집 pp.27-29, 1998

[3] 신진섭 "단어들의 연관성을 이용한 문서의 자동분류," 한국정보처리학회, 1999

[4] Joachims, T, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," Proc. of the 14th International Conference on Machine Learning ICML97, pp. 143-151, 1997

[5] 한승희, "문헌 클러스터링을 위한 유사계수간의 연관성 측정," 한국정보관리학회 학술대회 논문집 pp.25-28, 1999

[6] 최신정보검색론, 구미무역(주) 출판부, 1990

[7] 김호성 "용어 빈도수를 이용한 영문 문헌 정보의 점진적인 개념적 집단화," 한국정보과학회 논문지 Vol.19 pp.12-23, 1992

[8] 정영미 "지식 자동분류를 위한 유사성 척도의 비교 평가," 데이타베이스진흥센터 제 2회 디지털 도서관 컨퍼런스 논문집, pp.87-97, 1999

[9] 신진섭 "웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델," 박사학위 논문, 2000

[10] E. H. Han., et. al., "clustering based on association rule hypergraph," In Proc. of the Workshop on Research Issues on Data Mining and Knowledge Discoevery, pp.9-13, 1997

[11]. 서성보 " 주요 항목 집합을 이용한 문서 클러스터링 및 연관 탐사 기법," 한국정보과학회, pp.169-171, 2000

[12] 한국어 형태소 분석기-HAM(Hangul Analysis Module)," http://ham.hansung.ac.kr/ham/ham-intr.html

[13] 이재윤 "문헌 자동분류에서 용어 가중치 기법에 대한 연구," 한국정보관리학회 학술대회 논문집 pp.41-44, 2000

[14] Agrawal R, Strikant, "Fast algorithms for mining association rule," In Proc Of 20th intl. conf. On VLDB, pp.487-499, 1994